Bio-PDMS: A Structural Ontology-Driven Peer Data Management System (PDMS) for Biology

Tarczy-Hornoch, P.*, Brinkley, J., Shaker, R., Mork, P., Donelson, L., Barrier, M., Gennari, J., Rosse, C., Rossini, A., Sucio, D., Halevy, A.

Biomedical and Health Informatics, Biological Structure, Computer Science and Engineering, University of Washington, Seattle, WA, USA

For diverse groups of biology researchers, a common theme is the need to manage, share, integrate, and analyze heterogeneous, multi-scale, distributed data. For example, multiple University of Washington (UW) investigators and labs want to manage and share with one another their own diverse experimental data (e.g. gene expression, protein expression, and phenotypic manifestations at cellular, microscopic and gross levels of structure), analyze these data across UW investigators, and integrate these data with information in multiple public databases and knowledge bases. Bio-PDMS (Peer Data Management System) is an evolving architecture being developed to meet these needs.

The design and prototyping of Bio-PDMS is one part of the UW BISTI planning process to establish an Interdisciplinary Center for Structural Informatics. Bio-PDMS leverages the work of UW researchers in informatics and computer science in the areas of: a) knowledge base driven biological data integration (Tarczy-Hornoch, BioMediator, www.biomediator.org), b) open source software for analysis and comprehension of genomic data (Rossini, BioConductor, www.bioconductor.org), c) the UW Human Brain Project (Brinkley, UWHBP, sig.biostr.washington.edu), d) the Foundational Model of Anatomy (Rosse & Brinkley, FMA), sig.biostr.washington.edu), and e) peer data management systems (Halevy & Sucio, Piazza, data.cs.washington.edu/p2p/piazza/).

Current research at the UW independently addresses components of this broader vision of a biological data management system. The BioMediator biological data integration system provides a theoretical and practical foundation for data integration across diverse biological data sources via a knowledge base driven centralized federated database model. Piazza is a distributed peer data management system that provides a foundation for sharing of experimental data by peers via pairwise semantic mappings that allow data sources to cooperate without a central schema. The BioConductor project provides a basis for development and distribution of shared robust and reproducible analytic tools for statistically based genomic data analysis. The FMA ontology provides a common framework for organizing phenotypic manifestations. The UWHBP provides experience in structure-based biomedical experiment management systems.

The evolving Bio-PDMS model and architecture is being prototyped and refined by combining key elements of these existing systems. The BioMediator biological data integration system and the BioConductor genomic data analysis system have been combined in a gene expression array annotation system that provides a working prototype of a single system which combines data integration, annotation and analysis. The BioMediator system has been linked directly to flat file experimental protein expression data for purposes of annotating proteins of interest, demonstrating the ability to merge analysis of private experimental data with public databases and knowledge bases. The ability of the BioMediator mediated schema to access and use the Gene Ontology (GO) has demonstrated the feasibility and utility of including community-curated ontologies into a data management system. Ongoing work is focused on decentralizing the architecture into a peer model, adding complex spatial/temporal data types, aligning the FMA, GO, and other phenotype/genotype ontologies to provide a common semantic base.

Federal grant support: R01 HG02288, P20 LM007714, T15 LM07442, R01 DC02310